Squaring the Circle: More Generalizable Dialogue Discourse Parsing with Less Supervision

Anonymous ACL submission

Abstract

Discourse analysis plays a crucial role in Natural Language Processing, with discourse relation prediction arguably being the most difficult task in discourse parsing. Previous studies have generally focused on explicit or implicit discourse relation classification in monologues, leaving dialogue an under-explored domain. Facing the data scarcity issue, we propose to leverage self-training strategies based on Transformer backbone. Moreover, we design the first semi-supervised full discourse parsing pipeline that sequentially conducts parsing tasks. Using only 50 examples as gold training data, our relation prediction module achieves 58.4 in accuracy on the STAC corpus, close to supervised state-of-the-art. Full parsing results show notable improvements compared to the supervised models both in-domain (gaming) and cross-domain (technical chat), with better stability.

1 Introduction

Discourse analysis aims at uncovering the inherent structure of documents and has demonstrated its usefulness in various downstream applications, from sentiment analysis or fake news detection (Bhatia et al., 2015; Karimi and Tang, 2019), to summarization or machine translation (Chen and Yang, 2021; Chen et al., 2020). Existing research efforts have focused on automatically extracting discourse structures through tasks such as discourse relation prediction (Shi and Demberg, 2019; Wu et al., 2022) and discourse parsing (Joty et al., 2015; Kobayashi et al., 2020). The latter is usually conducted within the Rhetorical Structure Theory (RST) (Mann and Thompson, 1987) or the Segmented Discourse Representation Theory (SDRT) (Asher et al., 2003) where discourse structures are presented as trees or graphs. Automatic discourse parsing consists of extracting such structures from documents, where spans of text – known as Elementary Discourse Units (EDUs) – are linked by semantic-pragmatic relations such as *Explanation, Acknowledgment, Contrast, etc*.

Current data-driven methods for discourse parsing have predominantly been applied to monologues, leading to limited availability and generalizability of discourse parsers for dialogues. As dialogue data soared in all kinds of forms, such as online teaching and meetings, the need for automatic analysis systems has rapidly increased. However, one of the main hurdles in developing high-functioning parsing models is the scarcity of annotated data, along with limitations of supervised approaches in cross-domain scenarios (Liu and Chen, 2021). Strategic Conversations corpus (STAC) (Asher et al., 2016) – the most commonly used SDRT-annotated dialogue dataset – contains merely 1000 short documents. The labeling effort being expensive in terms of time and labor costs, it appears unlikely to create new large-scale expert-annotated datasets. Semi-supervised strategies are thus appealing. A few studies proposed weak or distant supervision for naked tree building (Badene et al., 2019; Li et al., 2023) while missing the important relation information. Remarkably, despite recent powerful Large Language Models (LLMs) such as ChatGPT excel in many NLP tasks, discourse parsing remains a significant challenge, given their poor performance (Chan et al., 2023a).

In this paper, we extend the bootstrapping approach to dialogues with even less annotated data, by relying on self-training (Yarowsky, 1995) where a model is used to produce pseudo labels and increase training data, a simple method shown as effective (Rosenberg et al., 2005). Using the BERT model (Devlin et al., 2019) as a base classifier and applying self-training, we achieve competitive results on a 16-way classification on STAC using only 50 dialogues for initial training. We also build a pipeline upon Li et al. (2023)'s work to perform full parsing, where we assign discourse relations on established structures, giving important extensions on semi-supervised approaches for dialogues until now limited to naked structures. Our pipeline yields 38.6 micro-F_1 score with gold EDUs and 32.8 with predicted EDUs: representing strong baselines for discourse parsing in dialogues with minimal supervision. This pipeline, or *structure-then-relation* approach, allows for a flexible architecture and greater generalizability. We further conduct cross-domain experiments by testing on a re-annotated subset of Molweni (Li et al., 2020) – a Ubuntu dataset. Despite the domain difference, our pipeline shows remarkable performances (link 75.6, link and relation 31.2), outperforming supervised SOTA models by a large margin[1].

To summarize our contributions in this work: we propose (1) a simple but effective method that requires minimal supervision for discourse relation prediction; (2) a flexible discourse parsing pipeline that handles all tasks in sequence and exhibits strong generalizability; (3) a comprehensive comparison with supervised models and in-depth exploration across in-domain and cross-domain scenarios; and (4) a small human-annotated discourse dataset in the technical chat domain which we will make public and support cross-domain evaluation.

2 Related Work

In recent years, there has been an increasing interest in discourse parsing in dialogues. Since the release of the STAC corpus, a range of discourse parsers has emerged, including classic statistical models (Afantenos et al., 2015; Perret et al., 2016) and neural architecture models (Shi and Huang, 2019; Wang et al., 2021; Chi and Rudnicky, 2022), some of which are trained within multi-task learning framework (Yang et al., 2021; Fan et al., 2022). Although these supervised models achieve good performance on STAC, they face limitations when applied to cross-domain scenarios (Liu and Chen, 2021). To address the challenge of data scarcity, researchers turn to weakly and semi-supervised methods, as done by Badene et al. (2019) and Li et al. (2023). For monologues, Nishida and Matsumoto (2022) show that co-training can considerably increase cross-domain performance, but they benefit from a larger amount of annotated data than we do for dialogues. Despite the revolutionary achievements offered by LLMs (Ouyang et al., 2022; Touvron et al., 2023), they remain notably behind fully and semi-supervised benchmarks in discourse parsing. Chan et al. (2023a) illustrate that ChatGPT struggles on STAC with 50% F_1 gap from supervised models. Fan and Jiang (2023) find that ChatGPT tends to establish discourse structures in a linear fashion. While in-context learning methods are helpful, their enhancement is limited.

On the other hand, discourse relation prediction as an individual task receives rich attention, mostly conducted on the Penn Discourse Treebank (PDTB) (Webber et al., 2019). This line of research can be categorized into explicit (Nie et al., 2019) and implicit relation identification (Rutherford et al., 2017). Semi-supervised models have been mostly limited to implicit ones either relying on synthetic data (Xu et al., 2018) or translations (Shi et al., 2019). These methods create pseudo-labeled data by using expert-composed rules or convenient linguistic resources: both in short in our case. The more recent effort seeks supervision from Pre-trained Language Models (PLMs) (Shi and Demberg, 2019; Arslan et al., 2021) as they show superior performance for many classification tasks. In the context of semi- and weakly supervised learning, PLMs have been used as reliable classifiers to produce pseudo labels (Meng et al., 2020; Yu et al., 2021). Through prompt adaptation, Chan et al. (2023b) reveal that implicit relation prediction is still a tricky task for ChatGPT, a finding that aligns with the results in discourse parsing.

3 Discourse Parsing Pipeline

A standard full discourse parsing involves three tasks: EDU segmentation, link attachment, and relation prediction (Figure 1). Most previous work applies a *structure-then-relation* approach (Afantenos et al., 2015; Shi and Huang, 2019; Liu and Chen, 2021). We follow the pipeline by providing relations on the established discourse structures.

3.1 Preliminary

Our work is founded on Li et al. (2023) which entails the extraction of discourse structures from the attention matrices in PLMs. In that work, the original BART model (Lewis et al., 2020) is fine-tuned with dialogue-tailored Sentence Ordering task to better encode dialogue structures. In each attention head, the attention values among EDUs can be seen as edge weights. Thus, by using a Maximum Spanning Tree algorithm, they obtain discourse tree structures. That work proves that with just 50 examples, the optimal attention head can be consis-

[1]Our code will be made available at URL.

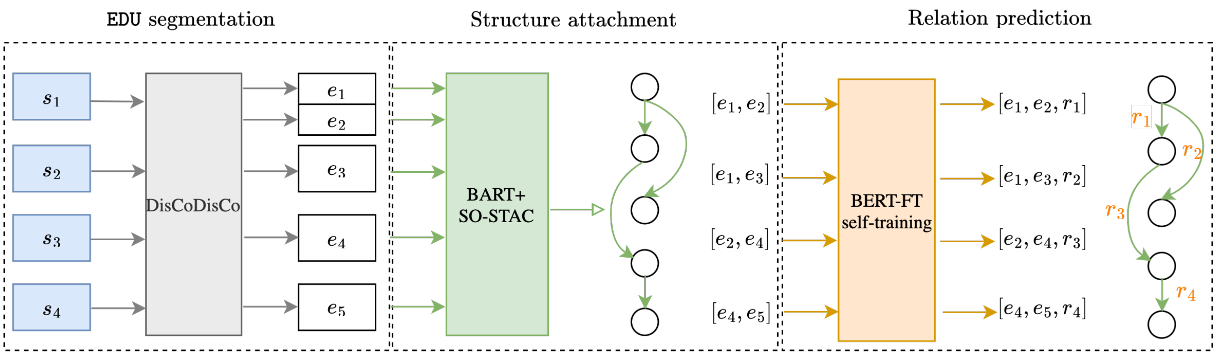

Figure 1: Semi-supervised discourse parsing pipeline proposition. s are utterances; e are EDUs; r are rhetorical relations. DisCoDisCo model is proposed in Gessler et al. (2021). BART+SO-STAC is BART model fine-tuned on Sentence Ordering task (Li et al., 2023). BERT-FT is BERT fine-tuned with self-training for relation prediction.

tently located. The extracted structures on STAC are found to be non-trivial, achieving 59.3 F_1 score.

Although most previous work begins with gold EDUs, we consider it crucial to evaluate in a deployed scenario where the parser performs EDU segmentation first. We thus integrate DisCoDisCo (Gessler et al., 2021), a straightforward sequence tagging model pre-trained on a random sample of 50 STAC dialogues, into the complete pipeline.

3.2 Relation Prediction Module

Following the setup in DISRPT shared tasks[2], we regard relation identification as multi-way classification where we classify every pair of head and dependent EDUs individually. EDU pairs reflect local coherence. A model trained in this setting is easily applicable to other discourse frameworks.

Self-Training: Our relation prediction module contains a classifier $\mathcal{M}$, a small amount of labeled data $\mathcal{L}$, and a large amount of unannotated data $\mathcal{U}$. The training process is as follow: $\mathcal{M}$ is trained on $\mathcal{L}$ to provide predictions (pseudo labels) on $\mathcal{U}$; then, under pre-defined selection criteria, a subset $\mathcal{S} \subset \mathcal{U}$ is sampled and merged with $\mathcal{L}$ for a new round of re-training. $\mathcal{M}$ can be re-trained for many rounds until a stopping criterion is met.

Classifier $\mathcal{M}$: Our classifier is an uncased BERT base model appended with a linear projection and softmax layer to produce relation probabilities. BERT has shown superior performance in discourse-related tasks (Chen et al., 2019; Atwell et al., 2021) and is the language backbone of current SOTA model for relation on STAC (Gessler et al., 2021). We prepare the input relation pairs by following the Next Sentence Prediction pattern as in Shi and Demberg (2019): a [CLS] token begins the sequence, followed by the first EDU, [SEP], and the second EDU. As additional feature, we only add the speaker marker at the beginning of the EDUs since it is the only feature we found decisive among the ones used in Gessler et al. (2021).[3]

Sample Selection Criteria: At each round, $\mathcal{M}$ gives pseudo labels on $\mathcal{U}$. The key challenges are how to measure the confidence of predictions and how to select a reliable subset $\mathcal{S}$. We loosely translate the output probabilities in $\mathcal{M}$ as its predictive confidence, enabling sorting predicted pairs. We then define two selection criteria inspired by Steedman et al. (2003); Du et al. (2021), either focusing on the confidence or combining it with class variety: (a) **Top-k**: select the top k pseudo-labeled data. k starts at 800 and increments up 7800, with an interval of 1000. This range corresponds to the top $N \times k'$ where $k' \in [0.0, 0.1]$ criterion in Nishida and Matsumoto (2022); (b) **Top-class-k**: select the most confident pseudo-labeled data in each class and together results in k examples. The label ratio is maintained between $\mathcal{L}$ and the augmented set $\mathcal{S}$. k has the same value as in Top-k.

4 Molweni Re-Annotation

To evaluate the cross-domain adaptability of our parsing pipeline, we release a newly annotated dataset, "Molweni-clean", sourced from the Molweni corpus (Li et al., 2020). Molweni contains 10,000 SDRT-annotated documents from the Ubuntu Chat Corpus (Lowe et al., 2015). However, it presents heavily redundant documents and

[2]https://github.com/disrpt/sharedtask2023/.

[3]Our supervised model gives 64.9 versus feature-enhanced DisCoDisCo 65.0 (Gessler et al., 2021).

	Avg branch	Avg depth	%leaf	Arc length
Molweni	1.63	6.0	0.39	0.23
∼-clean	1.29	6.8	0.28	0.19

Table 1: Tree properties in original Molweni test set and Molweni-clean. Arc length is normalized.

Dataset	#Doc train	dev	test	#Turn /doc	#Tok /doc	#Spk /doc	#Rel type
STAC	947	105	109	11.0	48.4	3.0	16
Molweni	9000	500	500	8.8	104.7	3.5	16
∼-clean	-	-	50	8.5	91.1	3.2	16

Table 2: STAC, Molweni, and Molweni-clean statistics: number of documents, averaged speech turns, tokens, and speakers per document (turn/doc, tok/doc, spk/doc).

inconsistent annotations (Li et al., 2023), making the results less reliable. Therefore, we revised the annotation of a subset of Molweni to ensure a more robust evaluation (test only).

4.1 Molweni-clean Construction

Molweni test set comprises 500 documents that can be grouped into 105 clusters. Each cluster consists of highly similar dialogues, with only one or two differing utterances (Li et al., 2023). As the first step of our re-annotation process, we extract a single document from each cluster, ensuring that the selected subset contains no duplicates.

The re-annotation is carried out by 3 Ph.D. students who are fluent in English, specialized in semantics and discourse and are familiar with SDRT. We pre-selected 105 documents from the test set with no duplicates as our annotation candidates. A set of 8 documents is used for training the annotators who then annotate 10 documents in common, and 20 more separately, leading to a final subset of 50 dialogues[4]. The inter-annotator agreement (Cohen's Kappa) is strong (80.6%) for link attachment and moderate (57.0%) for full structure, similar to the scores in STAC (Asher et al., 2016), with details in Appendix B.1.

4.2 Molweni-clean Statistics

Structural Difference: More adjacent links are presented in Molweni-clean (76% vs. 68%). Intuitively, these are simpler structures. The trees in Molweni-clean are "taller" and "thinner", namely, with smaller branch sizes and larger tree depths. On average, Molweni-clean trees are one step deeper than the originally annotated ones, as shown in Table 1. Additionally, we find 3 documents in the original annotation that contain multiple roots, resulting in *forest* structures instead of trees.

Relation Distribution: Although the class distribution appears to be alike in the two annotations (details in Appendix B.2), the partition between the same (intra-) and different (inter-) speakers differs greatly. In Molweni-clean, we observe a much higher percentage of intra-speaker relations (14.7% vs. 3.8%). Certain relations, like *Continuation* and *Elaboration* — which, according to the annotation guideline, should typically occur more frequently within the same speaker — show a contrasting distribution in the original annotation. We present a case study in Appendix B.3.

[4]These annotations are publicly available at URL.

5 Experimental Setup

Datasets: For the in-domain scenario (gaming), we utilize STAC, a corpus comprising of online conversations that occur during the *Settlers of Catan* game. It contains in total 12,679 relation pairs in 1161 documents. We follow the split in Shi and Huang (2019). We randomly select a small part (700 pairs from 50 documents) of the train set as labeled data $\mathcal{L}$ and the remaining examples as raw data $\mathcal{U}$. A subset from the development set (664 pairs from 50 documents) is used for validation. All 1128 pairs (109 documents) in the test set are reserved for testing. The relation distribution is highly unbalanced, see Appendix A. For the cross-domain scenario (gaming to technical chat), we use documents from STAC as the labeled training data, and the 50 Molweni-clean documents as testing data. Table 2 shows the statistics.

Evaluation Metrics: For the relation prediction module, we report accuracy. For the full parsing pipeline, we employ the traditional evaluation metrics, namely, the micro-averaged F_1 scores for unlabeled attachment (link), relation prediction (rel), and labeled attachment (full).

Full Parsing Baselines: We compare against the state-of-art parsing model Structured-Joint (SJ) (Chi and Rudnicky, 2022). Since we work with small-data setup, we also compare with a simpler graph-based Arc-Factored dependency parser (McDonald et al., 2005), by following the implementation in Nishida and Matsumoto (2022). Furthermore, to gain insights from the latest LLMs, we show results from ChatGPT[5] (gpt-3.5-turbo model) using zero-shot and few-shot in-context learning (Chan et al., 2023a).

Implementation Details: In the relation prediction module, we use the BERT model from Huggingface (Wolf et al., 2020) and fine-tune for 10 epochs with batch of size 2, learning rate at $2e-5$, AdamW optimizers with a weight decay at 0.01. For self-training, we give maximum 20 epochs with early stopping at 5, based on the performance on the validation set. We choose 5 groups of labeled examples for initial training and report average accuracy with the standard deviation. The full pipeline is trained using 50 random documents from STAC training set and is executed 10 times.

6 Relation Prediction Module

6.1 Self-Training Results

Results for relation prediction are presented in Table 3. As baselines, we report scores of majority class *Question answer pair* (*QA pair*), the original frozen BERT base model and the fine-tuned BERT, both trained with 700 gold pairs. Using this latter model as a starting point, we present results for self-training (second part of Table 3) using two sample selection criteria: top-k and top-class-k. Both selection strategies show improved performances with self-training. When $k = 5800$, both strategies achieve their best scores. This value echos the selection strategy rank-above-$k\prime$ with $k\prime = 0.6$ in Nishida and Matsumoto (2022). For top-k selection, when k is small ($k < 2800$), the number and variety of selected pseudo-labeled data are small, resulting in lower accuracy than BERT-ft. When k is relaxed, the coverage of different classes of data increases, and the performance hits the highest point at 58.1. The accuracy then decreases, probably due to the noise of inaccurate pseudo-labeled data. In comparison, the top-class-k strategy consistently brings improvement over the initial BERT-ft model. It also exhibits an upward trend as k increases, reaching its peak at the optimal value of 5800, followed by a slight decline.

With a significant amount of unlabelled data, the self-training process can be repeated multiple times. However, limited by the data size in STAC, we can only test iterative learning with few values, $k \in [800, 1800, 2800]$. We define a stopping criterion at 3 and proceed with top-class-k selection strategy. We observe (two rightmost columns) additional improvements compared to the first loop, reaching 58.4 at best. We speculate that the model is re-trained slowly (smaller amount of data), but steadily (more reliable examples). We anticipate a better performance with more in-domain raw data.

[5] https://openai.com/blog/chatgpt.

Majority class				27.1
BERT (base 700)				$40.1_{0.8}$
BERT-ft (base 700)				$56.6_{1.0}$
Self-training #Pair	Top-k loop1	Top-class-k loop1	loop2	loop3
+ 800	$54.1_{3.0}$	$57.7_{1.1}$	$55.9_{1.1}$	**$58.1_{1.2}$**
+ 1800	$53.6_{3.6}$	$57.3_{1.6}$	**$58.4_{1.0}$**	$57.4_{2.1}$
+ 2800	$55.7_{1.9}$	$57.6_{0.3}$	$57.5_{1.5}$	$58.1_{2.2}$
+ 3800	$56.6_{2.1}$	$57.6_{1.6}$	-	-
+ 4800	$56.8_{0.5}$	$57.8_{1.2}$	-	-
+ 5800	**$58.1_{0.8}$**	**$58.0_{0.7}$**	-	-
+ 6800	$57.8_{1.0}$	$57.9_{0.9}$	-	-
+ 7800	$57.8_{0.7}$	$57.0_{2.3}$	-	-

Table 3: Baselines and BERT-ft model self-training results with Top-k and Top-class-k selection criteria. Scores are avg accuracy over 5 runs with standard deviation. Best score per row (resp. per column) is underlined (resp. bold). - not applicable due to data limitation.

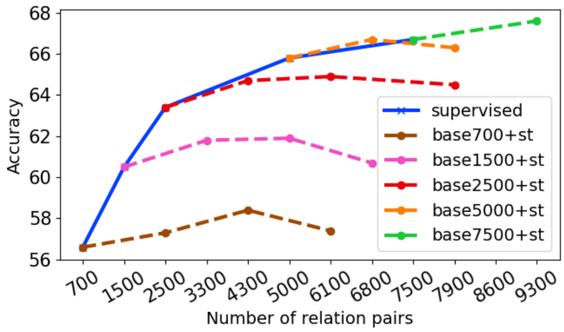

Figure 2: Accuracy of fully supervised model (solid line) and semi-supervised model with {700, 1500, 2500, 5000, 7500} base training data (dashed lines). x-axis: #relation pairs; y-axis: model accuracy on STAC.

6.2 Analysis: Model Calibration

One key challenge in self-training is to select error-free and high-coverage subsets from the pseudo-labeled data. Top-class-k selection considers the coverage aspect and less prone to overfitting. However, good coverage does not imply reliable prediction. The model could fall short in some classes and bring in noise. In this section, we study the correlation between the model's predicted probabilities and the probabilities of correctness, also known as the calibration property (Brier, 1950; Jiang et al., 2021). We start by showing this property of base BERT-ft model (details in Appendix C.1): frequent relations (e.g. *QA pair* and *Comment*) present positive correlation while infrequent ones (e.g. *Alternation* and *Correction*) do not and have lower confidence. This shows the advantage of top-class-k strategy by adding these less confident but reliable examples. However, it also implies that the base model is not well-calibrated. We investigate two factors that may influence the model's calibration: enhancing the classifier's accuracy by training on more base data and employing iterative training.

Base Model Accuracy: We experimentally observe that with more base training data, the model performance continuously increases (e.g.: from 700 to 2500, accuracy increases by 7%). In particular, we test different sizes of base data: $\{700, 1500, 2500, 5000, 7500\}$ of relation pairs and re-train the model using top-class-k ($k = 1800$) selection criterion. The results are displayed in Figure 2. With larger base volume, the gap between self-trained model and fully supervised model keeps decreasing. Interestingly, when the base data hits 5000, self-trained model achieves comparable performance as 7500 fully supervised model (66.7%), indicating that 5000 relation pairs ($\approx$ 350 documents) is a threshold where self-trained model surpasses its supervised counterpart.

Iterative Training: The concept of multi-loop self-training aims to enhance the model's performance by incorporating additional training examples for the *infrequent* classes, thereby mitigating the under-fitting issue. We investigate the correlation evolution with three loops for the less-frequent labels (details in Appendix C.2). Tellingly, the confidence scores for less and non-frequent relations such as *Alternation* and *Contrast* increase from $[0.2, 0.3]$ to $[0.7, 1.0]$, coupled with higher prediction accuracy (+ 20% $\sim$ 40%), as displayed in the confusion matrix in Figure 9.

7 Full Discourse Parsing

7.1 In-Domain Evaluation and Analysis

In-domain performance is evaluated on the STAC test set, with results in Table 4 (left part).

Baselines: We replicate the SOTA supervised model Structured-Joint (SJ) (Chi and Rudnicky, 2022) which uses RoBERTa-base model (Liu et al., 2019) as backbone and employs 3-dimension attention to encode links and relations jointly. SJ includes a dummy root in each document for training, but the link between this node and the first EDU is counted in the evaluation which artificially inflates the scores. We replicate SJ with 947 and 50 training data and evaluate with and without dummy root, the latter matching our own fairer evaluation setting. Table 4 shows our replicated scores without dummy root (detailed comparison in Appendix D). We also compare with a simpler dependency parser Arc-Factored (AF) (McDonald et al., 2005). AF parser finds the globally optimal dependency structure using dynamic programming which can be decoded using Maximum Spanning Tree algorithms such as Eisner (Eisner, 1996). Lastly, we report the performance of unsupervised LLM ChatGPT-3.5.

Parsing Results: Our pipeline consists of an EDU segmenter (Gessler et al., 2021), a link attachment module (Li et al., 2023) which we replicate the experiments and obtain predicted links, and a pre-trained relation prediction module outlined in Section 3.2. We sample 50 annotated documents for supervision along the pipeline. As expected, the supervised SJ model with 947 training examples gives the best scores. However, when the training size drops to 50, our pipeline exhibits better performance compared to SJ and AF in both link attachment (59.3% vs. 55.1%) and relation prediction (62.0% vs. 61.1%) tasks, bringing noteworthy improvement of resp. 5 and 14 points in full parsing, coupled with greater stability. As for GPT-3.5, both zero-shot and few-shot in-context learning perform abysmally, suggesting that ChatGPT still suffers from poor understanding of discourse structures and that we can not simply depend on powerful LLMs for this task (Chan et al., 2023a). Using predicted EDUs, our full parsing score drops nearly 6 points. A similar loss is also observed for end-to-end RST-style parsing in Nguyen et al. (2021).

Pipeline Error Analysis: We examine the relation composition in each task module: correct (orange) and wrong relation prediction (blue), and missing relations due to lack of link attachment (green) and false EDU segmentation (gray), as displayed in Figure 3. The results show that errors in link attachment account for 40.8%. Among the correctly attached pairs, 61% are assigned proper relations. Notably, relations such as *QA pair*, *Elaboration*, and *Acknowledgement* are accurately predicted, while less frequent relations such as *Result*,

Train / Test	Train	STAC/STAC				STAC/Molweni-clean			STAC/Molweni		
	#Doc	EDU	Link	Rel	Full	Link	Rel	Full	Link	Rel	Full
SJ	947	-	$70.7_{0.5}$	$77.3_{1.2}$	$54.6_{0.7}$	$61.5_{3.4}$	$59.5_{4.3}$	$36.6_{3.8}$	$49.8_{3.6}$	$57.5_{2.9}$	$28.9_{2.8}$
SJ	50	-	$55.1_{3.5}$	$61.1_{2.1}$	$33.6_{2.2}$	$51.1_{6.4}$	$33.6_{9.5}$	$17.2_{5.3}$	$42.9_{5.6}$	$35.2_{10.1}$	$15.3_{5.3}$
AF	50	-	$42.7_{2.8}$	$56.4_{2.5}$	$24.0_{1.0}$	$53.7_{2.1}$	$38.8_{2.9}$	$20.9_{1.1}$	$45.9_{1.5}$	$41.4_{1.0}$	$19.0_{0.7}$
GPT3.5$_{\text{few shot}}$	3	-	20.7	24.1	7.3	-	-	-	-	-	-
GPT3.5$_{\text{zero shot}}$	-	-	20.0	22.8	4.4	-	-	-	-	-	-
Ours (gold EDU)	50	-	$\mathbf{59.3}_{0.7}$	$\mathbf{62.0}_{1.1}$	$\mathbf{38.6}_{0.7}$	$\mathbf{75.6}_{0.7}$	$\mathbf{41.3}_{3.8}$	$\mathbf{31.2}_{2.9}$	$\mathbf{61.5}_{0.7}$	$\mathbf{42.8}_{2.9}$	$\mathbf{26.3}_{1.7}$
Ours (pred EDU)	50	94.8	$52.2_{0.4}$	$61.2_{1.6}$	$32.8_{0.9}$	∼	∼	∼	∼	∼	∼

Table 4: Left: in-domain parsing results (STAC/STAC) with supervised parsers Structured Joint (SJ) (2022) and Arc-Factored (AF) (2022), unsupervised model ChatGPT (GPT-3.5) with few-shot ($n = 3$) in-context learning and zero-shot (2023a), and our semi-supervised pipeline (with gold and predicted EDU). Right: cross-domain parsing results on Molweni-clean (STAC/Molweni-clean) and original Molweni (STAC/Molweni). Scores are average micro-F$_1$ over 10 runs. In 50 train setup, best scores are in bold. "-" not applicable. "∼" same as previous row.

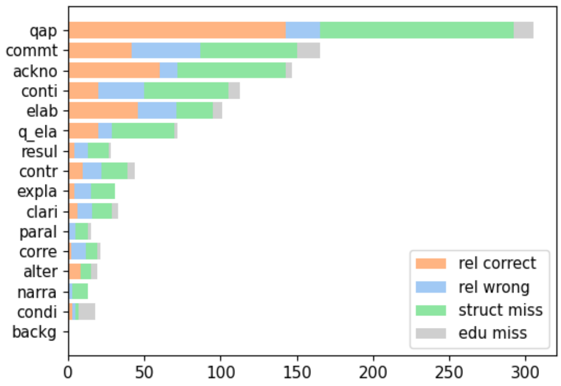

Figure 3: Full parsing result decomposition in relation prediction (orange and blue), link attachment (green), and EDU segmentation (grey). Numbers in Appendix E.

Explanation, and *Correction* require further improvements. We notice that the missing links often involve relation types that are accurately predicted (*QA pair* and *Acknowledgement*). This suggests that there is a high likelihood of accurately determining the discourse relations of connected pairs - a potential avenue for future improvement.

7.2 Cross-Domain Evaluation and Analysis

Cross-domain parsing is evaluated on the original Molweni test set and Molweni-clean, with SJ model and our pipeline trained on 50 STAC documents. Results are shown in Table 4 (right part).

Parsing Results: Our pipeline exhibits excellent performance on all tasks, outperforming the SJ model in terms of link (+24%), relation (+8%), and full parsing (+14%) on Molweni-clean dataset. Our pipeline for link attachment is particularly remarkable, surpassing even the fully trained SJ model (75.6 vs. 61.5). On relation prediction, SJ considers the tree structure and relation jointly, while our approach focuses on individual relation pairs. As texts across various genres demonstrate various structures, our approach, although more localized, is less influenced by the pre-existing structures, making it more suitable for general application. Furthermore, our model shows greater stability, whereas the SJ model is highly influenced by a particular domain. We notice similar behaviour on the original Molweni test set. Curiously, both SJ model and our pipeline exhibit improved performances on Molweni-clean, revealing the problem of inconsistencies in the initial annotation.

Molweni Cross-domain Annotation: We acknowledge that semi-supervised learning has an affinity for domain transfer. Taking one step further, we investigate automatic annotation on Molweni using STAC-trained model. The inconsistency of annotations in the original Molweni benefits this setup. We first de-duplicate repetitive documents in Molweni training and validation sets by taking one document per cluster (Sec. 4.1), which results in resp. 1865 and 107 documents. Trained on 50 STAC examples, our pipeline produces 1972 pseudo-labeled Molweni documents. These documents are used to train SJ in a supervised manner with the proposed hyper-parameters. In comparison, we also train the SJ model with Molweni's original annotation. Both models are evaluated on Molweni-clean, with results given in Table 6.

SJ model trained on pseudo-labeled Molweni gives better results on structure attachment (+9%) but under-performs its counterpart on relation prediction (-26%). Although the overall parsing score

Train / Test	Aug	STAC/STAC			STAC/Molweni-clean			STAC/Molweni		
	#Doc	Link	Rel	Full	Link	Rel	Full	Link	Rel	Full
SJ	-	$55.1_{3.5}$	$61.1_{2.1}$	$33.6_{2.2}$	$51.1_{6.4}$	$33.6_{9.5}$	$17.2_{5.3}$	$42.9_{5.6}$	$35.2_{10.1}$	$15.3_{5.3}$
SJ +self-train	50	$57.5_{2.2}$	$\mathbf{63.3_{1.4}}$	$36.4_{1.5}$	$51.6_{5.5}$	$34.3_{7.1}$	$17.6_{4.1}$	$42.9_{4.7}$	$34.5_{8.1}$	$14.8_{3.9}$
SJ +self-train	120	$57.2_{3.2}$	$62.7_{3.3}$	$35.9_{2.3}$	$54.3_{7.8}$	$40.3_{7.7}$	$21.9_{5.3}$	$45.7_{6.5}$	$39.2_{6.3}$	$18.0_{4.5}$
SJ +self-train	200	$57.4_{2.9}$	$63.1_{2.6}$	$36.2_{1.7}$	$56.4_{8.2}$	$38.4_{9.2}$	$21.8_{6.7}$	$46.6_{6.3}$	$38.7_{8.9}$	$18.1_{5.3}$
Ours	120	$\mathbf{59.3_{0.7}}$	$62.0_{1.1}$	$\mathbf{38.6_{0.7}}$	$\mathbf{75.6_{0.7}}$	$\mathbf{41.3_{3.8}}$	$\mathbf{31.2_{2.9}}$	$\mathbf{61.5_{0.7}}$	$\mathbf{42.8_{2.9}}$	$\mathbf{26.3_{1.7}}$

Table 5: Comparison between augmented SJ model (2022) (SJ +self-train) and ours in self-training setup across in-domain and cross-domain scenarios. SJ model is re-trained with the combination of 50 gold-standard data and {50, 120, 200} pseudo-labeled documents (Aug #doc). We show the best scores (average micro-F_1) in 3 loops.

Train on	#Doc	Link	Rel	Full
Molweni-pseudo	1865	$\mathbf{54.1_{0.6}}$	$56.3_{2.0}$	$30.6_{1.2}$
Molweni	1865	$45.7_{1.6}$	$\mathbf{82.7_{1.9}}$	$\mathbf{37.8_{1.1}}$

Table 6: SJ parsing results on Molweni-clean, trained on auto-annotated and original Molweni (resp. Molweni-pseudo, Molweni). Scores are average micro-F_1.

is inferior, the naked discourse structures in auto-annotated Molweni (Molweni-pseudo) are of better quality. This is encouraging, especially in the difficult cross-domain setup. As previous studies have shown, discourse structures alone are valuable features and can be employed in some downstream applications (Louis et al., 2010; Jia et al., 2020).

7.3 Self-Training the SJ Model

To understand the effectiveness of our relation prediction module, we conduct ablation studies by comparing our pipeline and SJ model with similar data volume, namely, we augment SJ model with self-training. Results are given in Table 5.

For the data augmentation, we select the pseudo-labeled documents with the highest average confidence scores, i.e., the average of predictive probabilities over all link and relation decisions in a document. Previous analysis (Sec. 6.2) shows that iterative training is beneficial, so we re-train SJ in a total of 3 loops. We test different sizes of augmentation data: {50, 120, 200} documents which correspond to resp. {800, 1800, 2800} relation pairs in our case. Over 3 loops, the largest augmentation attains 600 documents ($\approx$ 8000 relation pairs). It is important to note that although the SJ model jointly predicts structure and relation, our augmentation technique only focuses on relation prediction. Therefore, the augmentation would provide the SJ model with more structured supervision. Furthermore, our approach operates on a narrower scope, concentrating on relation pairs rather than entire conversations. In contrast, the SJ model's data augmentation is done at the document level. Hence, the comparison between our augmented model and the augmented SJ model would only be similar in terms of data volume, but not necessarily in terms of identical examples.

Given extra training data, SJ surpasses its base version in both in-domain (full +3%) and cross-domain (full +4%) contexts, with similar improvement in link attachment and relation prediction. This emphasizes the advantages of our self-training approach, apt for both basic and complex models. However, with the same augmented data size, the SJ model lags behind our pipeline, showcasing a 3 points difference in-domain and a sizable 10 points gap cross-domain, further attesting to the effectiveness of our simple approach.

8 Conclusion

In this study, we introduce a substantial extension to semi-supervised discourse parsing in dialogues by enhancing relation prediction via a self-training approach based on simple yet effective sample selection strategies. With a minimal training set of 50 examples, we produce highly competitive results that could be further improved with more in-domain raw data. Importantly, the efficacy of our discourse parsing pipeline is demonstrated across in-domain and cross-domain settings. We also contribute a small gold-standard discourse-enriched dialogue dataset, along with semi-supervised benchmarks for subsequent comparisons. Future work should explore the use of more out-of-domain raw data and investigate bootstrapping methods for relation prediction, while also improving on structure prediction, possibly with the same strategies.

Limitations

Following DISRPT shared task, we focused on individual EDU pair relation prediction for general application. This setting captures local coherence in dialogues and has shown great generalizability in cross-domain experiments. We based our work on a semi-supervised link attachment module and predicted relations only for linked EDU pairs. Showing effective, there is potential for further improvement in attachment performance by considering (high confident) predicted relations for unattached EDU pairs. By extending the self-training strategy to include link attachment, we could enhance the overall parsing performance and achieve better results in full parsing.

Facing the data sparsity issue, we utilized all relation pairs in STAC for self-training. However, we only tested small sizes of k in the iterative training due to the limited size of STAC. With more data, we should explore the re-training outcomes with larger values of k. It is thus intriguing to expand the set of un-annotated relations by considering out-of-domain data, obtained for instance from weak supervision (Sileo et al., 2019), or from monologues such as PDTB (Prasad et al., 2008).

Ethics Statement

We carefully selected the corpora to work with to mitigate any potential hateful and biased language. Before the re-annotation process, we provided instructions to the annotators, emphasizing the importance of being vigilant for any biased or insulting language in the data. In the event of encountering such language, they were instructed to immediately cease annotation and report the issue. Throughout the re-annotation of all 77 dialogues, no instances of inappropriate language were found. We have confidence that these dialogues are free from harmful content that may insult the annotators.

All the annotators are PhD students. They did not receive any specific compensation for their work on annotation. We recorded the time taken for the re-annotation process, which consisted of an initial training period of 3 hours followed by an average of 1.5 hour for every 10 dialogues. All annotation work was conducted during regular working hours. The annotators are free to utilize the annotations and any discourse-related content in this project for their studies.

References

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 928–937, Lisbon, Portugal. Association for Computational Linguistics.

Yusuf Arslan, Kevin Allix, Lisa Veiber, Cedric Lothritz, Tegawendé F Bissyandé, Jacques Klein, and Anne Goujon. 2021. A comparison of pre-trained language models for multi-class text classification in the financial domain. In *Companion Proceedings of the Web Conference 2021*, pages 260–268.

Nicholas Asher, Nicholas Michael Asher, and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2721–2727, Portorož, Slovenia. European Language Resources Association (ELRA).

Katherine Atwell, Junyi Jessy Li, and Malihe Alikhani. 2021. Where are we in discourse relation recognition? In *Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 314–325.

Sonia Badene, Kate Thompson, Jean-Pierre Lorré, and Nicholas Asher. 2019. Data programming for learning discourse structure. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 640–645, Florence, Italy. Association for Computational Linguistics.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon, Portugal. Association for Computational Linguistics.

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Chunkit Chan, Jiayang Cheng, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2023a. Chatgpt evaluation on sentence level relations: A focus on temporal, causal, and discourse relations. *arXiv preprint arXiv:2304.14827*.

Chunkit Chan, Xin Liu, Jiayang Cheng, Zihan Li, Yangqiu Song, Ginny Y Wong, and Simon See. 2023b. Discoprompt: Path prediction prompt tuning for implicit discourse relation recognition. *arXiv preprint arXiv:2305.03973*.

Jiaao Chen and Diyi Yang. 2021. Structure-aware abstractive conversation summarization via discourse and action graphs. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1380–1391, Online. Association for Computational Linguistics.

Junxuan Chen, Xiang Li, Jiarui Zhang, Chulun Zhou, Jianwei Cui, Bin Wang, and Jinsong Su. 2020. Modeling discourse structure for document-level neural machine translation. In *Proceedings of the First Workshop on Automatic Simultaneous Translation*, pages 30–36.

Mingda Chen, Zewei Chu, and Kevin Gimpel. 2019. Evaluation benchmarks and learning criteria for discourse-aware sentence representations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 649–662.

Ta-Chung Chi and Alexander Rudnicky. 2022. Structured dialogue discourse parsing. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 325–335.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jingfei Du, Édouard Grave, Beliz Gunel, Vishrav Chaudhary, Onur Celebi, Michael Auli, Veselin Stoyanov, and Alexis Conneau. 2021. Self-training improves pre-training for natural language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5408–5418.

Jason Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*.

Yaxin Fan and Feng Jiang. 2023. Uncovering the potential of chatgpt for discourse analysis in dialogue: An empirical study. *arXiv preprint arXiv:2305.08391*.

Yaxin Fan, Peifeng Li, Fang Kong, and Qiaoming Zhu. 2022. A distance-aware multi-task framework for conversational discourse parsing. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 912–921.

Luke Gessler, Shabnam Behzad, Yang Janet Liu, Siyao Peng, Yilun Zhu, and Amir Zeldes. 2021. Discodisco at the disrpt2021 shared task: A system for discourse segmentation, classification, and connective detection. In *Proceedings of the 2nd Shared Task on Discourse Relation Parsing and Treebanking (DISRPT 2021)*, pages 51–62.

Qi Jia, Yizhu Liu, Siyu Ren, Kenny Zhu, and Haifeng Tang. 2020. Multi-turn response selection using dialogue dependency relations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1911–1920.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Shafiq Joty, Giuseppe Carenini, and Raymond T Ng. 2015. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435.

Hamid Karimi and Jiliang Tang. 2019. Learning hierarchical discourse-level structure for fake news detection. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3432–3442.

Naoki Kobayashi, Tsutomu Hirao, Hidetaka Kamigaito, Manabu Okumura, and Masaaki Nagata. 2020. Top-down rst parsing utilizing granularity levels in documents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8099–8106.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chuyuan Li, Patrick Huber, Wen Xiao, Maxime Amblard, Chloé Braud, and Giuseppe Carenini. 2023. Discourse structure extraction from pre-trained and fine-tuned language models in dialogues. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 2517–2534.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2642–2652, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Zhengyuan Liu and Nancy Chen. 2021. Improving multi-party dialogue discourse parsing via domain integration. In *Proceedings of the 2nd Workshop on Computational Approaches to Discourse*, pages 122–127, Punta Cana, Dominican Republic and Online. Association for Computational Linguistics.

Annie Louis, Aravind K Joshi, and Ani Nenkova. 2010. Discourse indicators for content selection in summaization.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: Description and construction of text structures*. Springer.

Ryan McDonald, Fernando Pereira, Kiril Ribarov, and Jan Hajic. 2005. Non-projective dependency parsing using spanning tree algorithms. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 523–530.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text classification using label names only: A language model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9006–9017.

Thanh-Tung Nguyen, Xuan-Phi Nguyen, Shafiq Joty, and Xiaoli Li. 2021. Rst parsing from scratch. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1613–1625.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. Dissent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510.

Noriki Nishida and Yuji Matsumoto. 2022. Out-of-domain discourse dependency parsing via bootstrapping: An empirical analysis on its effectiveness and limitation. *Transactions of the Association for Computational Linguistics*, 10:127–144.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Jérémy Perret, Stergos Afantenos, Nicholas Asher, and Mathieu Morey. 2016. Integer linear programming for discourse parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 99–109, San Diego, California. Association for Computational Linguistics.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind Joshi, and Bonnie Webber. 2008. The penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.

Chuck Rosenberg, Martial Hebert, and Henry Schneiderman. 2005. Semi-supervised self-training of object detection models.

Attapol Rutherford, Vera Demberg, and Nianwen Xue. 2017. A systematic study of neural discourse models for implicit discourse relation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 281–291.

Wei Shi and Vera Demberg. 2019. Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5790–5796.

Wei Shi, Frances Yung, and Vera Demberg. 2019. Acquiring annotated data with cross-lingual explicitation for implicit discourse relation classification. *NAACL HLT 2019*, page 12.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7007–7014.

Damien Sileo, Tim Van De Cruys, Camille Pradel, and Philippe Muller. 2019. Mining discourse markers for unsupervised sentence representation learning. In *Proceedings of NAACL-HLT*, pages 3477–3486.

Mark Steedman, Rebecca Hwa, Stephen Clark, Miles Osborne, Anoop Sarkar, Julia Hockenmaier, Paul Ruhlen, Steven Baker, and Jeremiah Crim. 2003. Example selection for bootstrapping statistical parsers. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 236–243.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti

Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ante Wang, Linfeng Song, Hui Jiang, Shaopeng Lai, Junfeng Yao, Min Zhang, and Jinsong Su. 2021. A structure self-aware model for discourse parsing on multi-party dialogues. In *Proceedings of the Thirtieth International Conference on International Joint Conferences on Artificial Intelligence*.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Changxing Wu, Liuwen Cao, Yubin Ge, Yang Liu, Min Zhang, and Jinsong Su. 2022. A label dependence-aware sequence generation model for multi-level implicit discourse relation recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11486–11494.

Yang Xu, Yu Hong, Huibin Ruan, Jianmin Yao, Min Zhang, and Guodong Zhou. 2018. Using active learning to expand training data for implicit discourse relation recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 725–731.

Jingxuan Yang, Kerui Xu, Jun Xu, Si Li, Sheng Gao, Jun Guo, Nianwen Xue, and Ji-Rong Wen. 2021. A joint model for dropped pronoun recovery and conversational discourse parsing in chinese conversational speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1752–1763.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *33rd annual meeting of the association for computational linguistics*, pages 189–196.

Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. 2021. Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1063–1077.

A Class Distribution in STAC Corpus

See Table 7 for the relation distribution in train, development, and test sets in STAC.

Relation	Labeled train #	Labeled train %	Validation #	Validation %	Test #	Test %
QA pair	175	25.0	152	22.89	305	27.04
Comment	108	15.43	110	16.57	165	14.63
Ack	86	12.29	87	13.1	148	13.12
Continuation	65	9.29	69	10.39	113	10.02
Elaboration	64	9.14	52	7.83	101	8.95
Q-elab	36	5.14	30	4.52	72	6.38
Result	26	3.71	29	4.37	29	2.57
Contrast	32	4.57	29	4.37	44	3.9
Explanation	34	4.86	31	4.67	31	2.75
Clarif-Q	23	3.29	20	3.01	33	2.93
Parallel	10	1.43	14	2.11	15	1.33
Correction	12	1.71	11	1.66	21	1.86
Alternation	5	0.71	8	1.2	19	1.68
Narration	8	1.14	7	1.05	13	1.15
Conditional	12	1.71	10	1.51	18	1.6
Background	4	0.57	5	0.75	1	0.09
Total	700	100.0	664	100.0	1,128	100.0

Table 7: Rhetorical relations and frequencies in train subset, validation subset, and test sets in STAC. QA pair: question answer pair; Ack: acknowledgement; Q-elab: question elaboration; clarif-Q: clarification question.

B Molweni-clean Case Study

B.1 Inter-Annotator Agreement Detail

We calculate inter-annotator agreement scores on the 10 common documents using Cohen's Kappa metric from Scikit-learn library (Pedregosa et al., 2011). The results are given in Table 8. Our final subset contains 50 documents. Annotator 1 and 3 (R1 and R3) have the highest agreement scores, so we include their individual annotations (a total of 39 documents). We also take the 8 training examples where all the annotators have aligned annotations and 3 documents from annotator 2.

	Link	Link&Rel
R1-R2	79.3	51.8
R1-R3	80.6	57.0
R2-R3	76.6	54.3

Table 8: Cohen's Kappa inter-annotator agreement scores. R1, R2, R3 represent resp. annotator 1, 2, and 3.

B.2 Relation Distribution Comparison

See Table 9 for relation distribution in original Molweni subset and Molweni-clean. We show the same 50 documents for a fair comparison. More precisely, we decompose each relation into intra- and inter- speaker categories to refer the relation within the same and different speakers, respectively. Note that the difference in the total number of relations (370 vs 373) is due to the incomplete annotation in the original annotation of documents 7048, 8018, and 9042 where one document contains multiple roots, i.e., some nodes miss an incoming edge.

Relation	Molweni test #	Molweni test %intra	Molweni test %inter	Molweni-clean #	Molweni-clean % intra	Molweni-clean %inter
Comment	99	2.0	98.0	104	2.9	97.1
Clarif-Q	89	0	100	84	2.4	97.6
QA pair	86	0	100	91	1.1	98.9
Continuation	28	17.9	82.1	27	92.6	7.4
Q-elab	11	9.1	90.9	18	22.2	77.8
Result	11	0	100	10	20.0	80.0
Explanation	9	11.1	88.9	5	40.0	60.0
Ack	7	0	100	6	0	100
Elaboration	7	42.9	57.1	14	85.7	14.3
Narration	7	0	100	1	100	0
Conditional	5	20.0	80.0	2	0	100
Contrast	3	0	100	2	50.0	50.0
Correction	3	0	100	6	16.7	83.3
Background	3	0	100	2	0	100
Parallel	2	50.0	50.0	0	0	0
Alternation	0	0	0	1	100	0
Total	370	3.8	96.2	373	14.7	85.3

Table 9: Relations distribution in original Molweni test subset and Molweni-clean.

B.3 Case Study

We present a comparison of the original annotation and our revised version for document #1035, as shown in Figure 4 and 5, respectively. This dialogue happens between two speakers: cr1mson (short in C) and APT-GET_INSTALL_ (short in A). C is asking A about the "apt" command. We show the number of speech turn after the speaker marker. Speech turns start from 0:

C0: *apt-get i doubt my apt thing is bad though , i just installed ubuntu today*

A1: *wait ! i found a much easier way*

A2: *well , i want you to read all of that*

A3: *before you start mucking around in system files*

C4: *there was only a couple lines in it*

C5: *most of it was rem 'd out*

A6: *you are going to learn what all of them all from the url i just pasted*

C7: *i can always use more than one terminal*

C8: *okay , so i have to add or change a 'repository'*

The main difference is in the annotation of *Complex Discourse Units* (CDUs) – several EDUs group together to form a common rhetorical function (Asher et al., 2016). In this example, the first CDU consists of three speech turns (A1, A2, A3) where A2 and A3 elaborate A1 by presenting a "much easier way". Between A2 and A3 it is a continuation. We can write as *Elaboration*(A1, *Continuation*(A2, A3)). This is a similar case with the example (58) in STAC annotation manual[6]. The original annotation, on the other hand, does not capture the accurate inner-CDU relations and roughly attaches every EDU inside the CDU with the first utterance C0.

Another CDU contains the speech turns C4 and C5. C5 continues C4 and together they provide a comment to A. Furthermore, we believe that CDU (C4, C5) should be linked to A2 instead of A3 since A2 and A3 are attached with a subordinating conjunction marker "before", which makes A3 *head* of this CDU. Semantically, "only a couple lines" also echos with "all of that". However, the original annotation does not capture the relationship between C4 and C5 and only link them individually to the previous utterance A3.

For each training document, annotators went through a similar discussion in order to reach consensus on difficult or ambiguous cases. We believe that this stage contributes to our improved understanding of dialogue content and the SDRT framework, and facilitate the production of more reliable annotations.

Figure 4: Original annotation of document 1035.

Figure 5: Re-annotated structure of document 1035.

C Class-wise Correlation Between Confidence and Accuracy

C.1 Correlation with Base Model

We investigate the correlation between class-wise confidence scores and prediction accuracy. For better readability, we divide 16 relations into 3 groups based on their frequency in the STAC corpus, as shown from top to bottom in the Figure 6. Recall

[6]https://www.irit.fr/STAC/stac-annotation-manual.pdf.

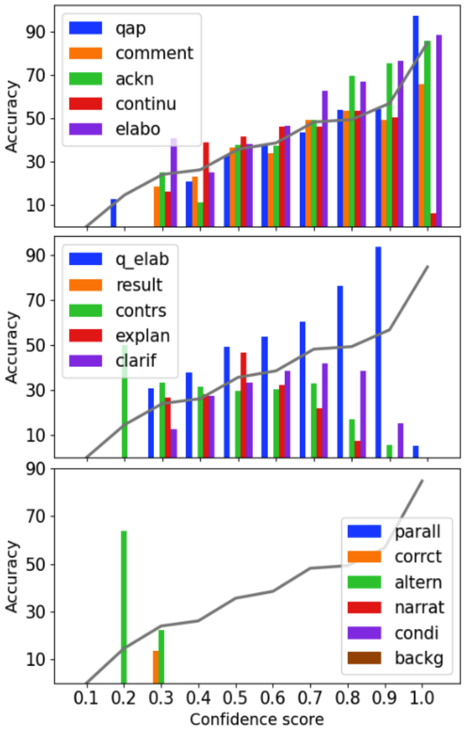

Figure 6: Relation class-wise accuracy and confidence score correlation in the base BERT-ft model. From top to bottom: the 5 most frequent, 5 medium-frequent, and 6 *infrequent* classes. The gray line is the aggregated score of all 16 relations.

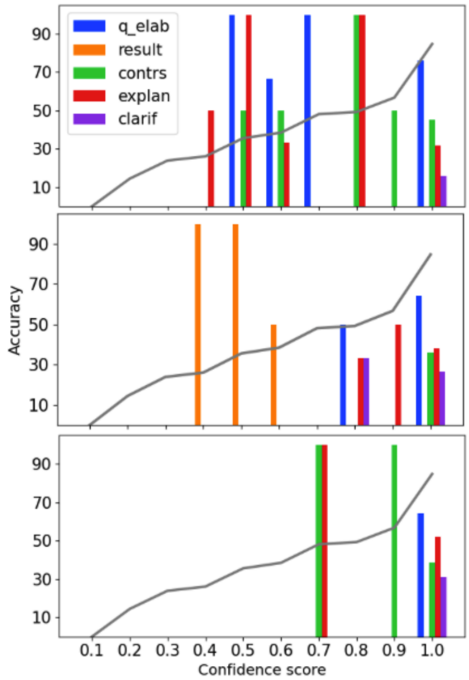

Figure 7: Accuracy and confidence score of the five medium-frequent relations in loop $\{1, 2, 3\}$.

that we translate confidence score with model's prediction probability.

The top plot in Figure 6 shows the first 5 relations: *QAP*, *Comment*, *Acknowledgement*, *Continuation*, and *Elaboration*. They are the most frequent relations. They show good positive correlation between the confidence and accuracy.

The middle plot in Figure 6 shows 5 medium-frequent relations: *Question elaboration*, *Result*, *Contrast*, *Explanation*, and *Clarification*. These relations have a frequency less than 10% and higher than 2% in STAC. The density of the bars moves towards the center compared to that with frequent relations, suggesting that the model is less *confident* to give predictions for these relations.

Finally, the last group contains six *infrequent* relations, as shown in bottom in Figure 6. They are the least present and the most difficult to predict. From this plot, we see that *Parallel*, *Narration*, *Conditional*, and *Background* are completely missing, while *Alternative* and *Correction* are correctly predicted with rather low confidence ($\in [0.2, 0.3]$).

C.2 Iterative Self-training Enhance Correlation for *Infrequent* Classes

Figure 7 and Figure 8 shows the changes of correlation during three loops. During iterative training, we observe that medium and the least frequent labels typically gain better correlation between accuracy and confidence scores, demonstrating that iterative training is good reinforcement for *infrequent* classes.

This observation is further proved in the confusion matrices, as displayed in Figure 9. A clear observation is that the *infrequent* classes has some recall improvement along self-training, typically for *Correction* and *Alternation*. For medium-frequent classes, *Result*, *Contrast*, and *Explanation* also obtain higher recall.

D SJ Model Reproduction Experiments

Table 10 shows the reproduction results on SJ model. Tellingly, removing the dummy roots leads to a noticeable drop, from around 59 to 54.6 in full parsing, which is even larger (-8 points) in cross-domain setting.

E Full Parsing Result Decomposition

Table 11 reports scores per class in each step of discourse parsing.

Train / Test		STAC/STAC			STAC/Molweni-clean			STAC/Molweni		
	#Train	Link	Rel	Link&Rel	Link	Rel	Link&Rel	Link	Rel	Link&Rel
(1) SJ reported scores	947	74.4	-	59.6	-	-	-	64.5	-	38.0
(2) SJ w dummy	947	$73.4_{0.4}$	$80.1_{1.1}$	$58.8_{0.7}$	$66.0_{3.0}$	$66.8_{3.5}$	$44.1_{3.3}$	$55.2_{3.1}$	$66.2_{2.7}$	$36.9_{2.4}$
(3) SJ w/o dummy	947	$70.7_{0.5}$	$77.3_{1.2}$	$54.6_{0.7}$	$61.5_{3.4}$	$59.5_{4.3}$	$36.6_{3.8}$	$49.8_{3.6}$	$57.5_{2.9}$	$28.9_{2.8}$
(4) SJ w dummy	50	$58.6_{2.7}$	$66.8_{1.8}$	$38.9_{1.9}$	$56.8_{5.6}$	$47.6_{7.5}$	$27.0_{4.7}$	$49.3_{5.0}$	$50.2_{7.1}$	$24.9_{4.7}$
(5) SJ w/o dummy	50	$55.1_{3.5}$	$61.1_{2.1}$	$33.6_{2.2}$	$51.1_{6.4}$	$33.6_{9.5}$	$17.2_{5.3}$	$42.9_{5.6}$	$35.2_{10.1}$	$15.3_{5.3}$

Table 10: SJ model reproduction (row 2-5) in different setups: in-domain and cross-domain, with different train sizes, and with or without dummy root. Scores are average F_1 over 10 runs. First row from the paper (2022).

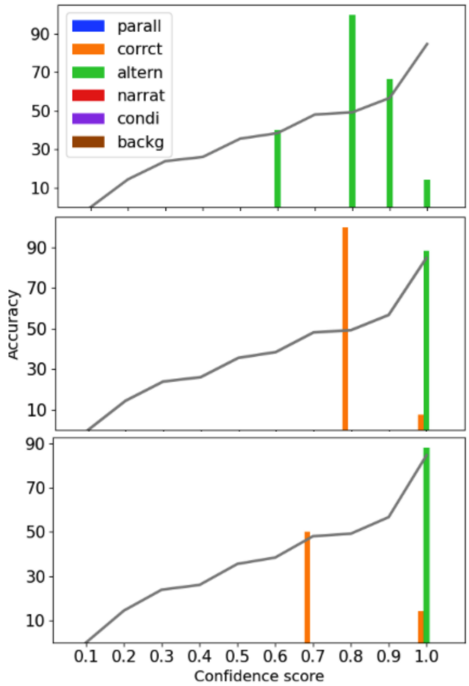

Figure 8: *Infrequent* relation accuracy and confidence scores, loop {1, 2, 3}.

Relation	#(%) correct	#(%) False relation	#(%) False link	#(%) False EDU
qap	143 (**46.9**)	22 (7.2)	127 (41.6)	13 (4.3)
commt	42 (25.5)	45 (27.3)	63 (38.2)	15 (9.1)
ackno	60 (40.5)	13 (8.8)	71 (48.0)	4 (2.7)
conti	20 (17.7)	30 (26.5)	55 (48.7)	8 (7.1)
elab	46 (**45.5**)	25 (24.8)	24 (23.8)	6 (5.9)
q_ela	20 (27.8)	9 (12.5)	41 (**57.0**)	2 (2.8)
resul	5 (17.2)	9 (**31.0**)	14 (48.3)	1 (3.5)
contr	10 (22.7)	12 (27.3)	17 (38.6)	5 (11.4)
expla	4 (12.9)	11 (**35.5**)	16 (51.6)	0 (0)
clari	6 (18.2)	10 (30.3)	13 (39.4)	4 (12.1)
paral	1 (6.7)	4 (26.7)	8 (**53.3**)	2 (**13.3**)
corre	2 (9.5)	10 (**47.6**)	7 (33.3)	2 (9.5)
alter	8 (**42.1**)	0 (0)	7 (36.8)	4 (**21.1**)
narra	0 (0)	3 (23.1)	10 (**76.9**)	0 (0)
condi	3 (16.7)	2 (11.1)	2 (11.1)	11 (**61.1**)
backg	0 (0)	0 (0)	1 (100)	0 (0)
Total	370 (32.8)	205 (18.2)	476 (42.2)	77 (6.8)

Table 11: Class-wise performance on relation prediction, link attachment, and EDU segmentation modules.

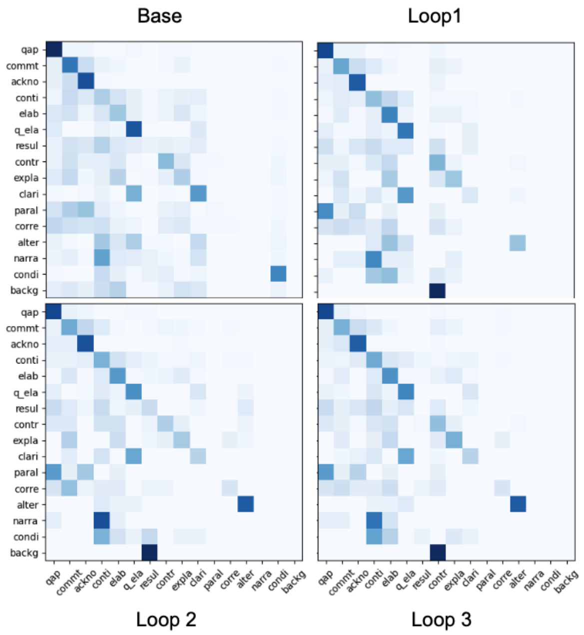

Figure 9: Confusion matrices in the base model and self-trained model with multiple loops. Relations (top to bottom, left to right): *QA pair, comment, acknowledgement, continuation, elaboration, question elaboration, result, contrast, explanation, clarification question, parallel, correction, alternation, narration, conditional, background.*